IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 5, 2019, accepted December 23, 2019, date of publication January 7, 2020, date of current version January 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2964540

Attention Mask R-CNN for Ship Detection and Segmentation From Remote Sensing Images

XUAN NIE[1], MENGYANG DUAN[1], HAOXUAN DING[2], BINGLIANG HU[3], AND EDWARD K. WONG[4]

[1]School of Software, Northwestern Polytechnical University, Xi'an 710072, China
[2]School of Power and Energy, Northwestern Polytechnical University, Xi'an 710072, China
[3]Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China
[4]NYU Tandon School of Engineering, Brooklyn, NY 11201, USA

Corresponding author: Xuan Nie (xnie@nwpu.edu.cn)

This work was supported in part by the 2018 Science and Technology Plan of Shaanxi Province under Project 2018ZDXM-GY-108, and in part by the Open-End Funds of Key Laboratory of Spectral Imaging Technology, CAS, under Project LSIT201706D.

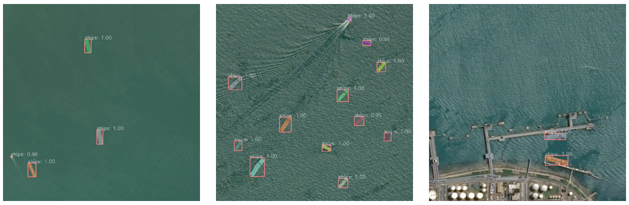
ABSTRACT In recent years, ship detection in satellite remote sensing images has become an important research topic. Most existing methods detect ships by using a rectangular bounding box but do not perform segmentation down to the pixel level. This paper proposes a ship detection and segmentation method based on an improved Mask R-CNN model. Our proposed method can accurately detect and segment ships at the pixel level. By adding a bottom-up structure to the FPN structure of Mask R-CNN, the path between the lower layers and the topmost layer is shortened, allowing the lower layer features to be more effectively utilized at the top layer. In the bottom-up structure, we use channel-wise attention to assign weights in each channel and use the spatial attention mechanism to assign a corresponding weight at each pixel in the feature maps. This allows the feature maps to respond better to the target's features. Using our method, the detection and segmentation mAPs increased from 70.6% and 62.0% to 76.1% and 65.8%, respectively.

INDEX TERMS Computer vision, object detection, object segmentation, remote sensing.

I. INTRODUCTION

Nowadays, the marine transportation industry is making advances at a very fast pace. The rapid growth in the number of ships and shipping volume have also caused an increase in the number of maritime violations. Automated ship detection can help to obtain ship distribution information. It plays an increasingly important role in maritime surveillance, monitoring and traffic supervision. It can help to control illegal fishing and cargo transportation. In recent years, ship detection in satellite remote sensing images has become an important research topic. For example, [1] used structured forest edge detection, morphological image processing and support vector machine (SVM) to detect ships from satellite images downloaded from Google Earth. Synthetic Aperture Radar (SAR), which allows imaging during both day time and night time, has attracted the attention of many researchers. There has been considerable amount of prior work in SAR image ship detection. Many ship detection methods in SAR images are based on CFAR [2]–[4]. In recent years, CNN (convolutional neural network) has become the dominant method for image classification, target detection and segmentation [5]–[7]. Many detection methods based on deep learning have been developed and have achieved good results [8]–[13]. Deep CNN networks, such as Faster R-CNN [10], YOLO [12] and SSD [13], can detect, localize and predict the label of the target. In recent years, object detection in remote sensing satellite images using deep learning methods has become a hot research topic [14]–[16] and it has become a trend to detect ships in satellite remote sensing images by using deep learning methods [17]–[26]. CNN-based methods have been applied to ship detection in SAR images [21] and for detecting land targets [22]. They have achieved better performance than traditional methods. Reference [17] adopts Faster R-CNN structure, which fuses deep semantic and shallow high-resolution features in both the RPN and Region of Interest (RoI) layers, and improved the detection accuracy of small ships. Kang *et al.* [18] used Faster R-CNN for detection and employed CFAR to pick up small targets.

The associate editor coordinating the review of this manuscript and approving it for publication was Gianluigi Ciocca.



(a) (b) (c)

FIGURE 1. Detection and segmentation results using our model. (a) Sea area 1, (b) sea area 2, (c) harbor area.

In summary, previous studies have shown that CNN-based methods can detect ocean targets more accurately than CFAR and other feature-based methods. However, the above methods detect ships by using a bounding box and do not perform ship segmentation down to pixel level.

Mask R-CNN [27] is a convolutional neural network based on Faster R-CNN [10]. This neural network can detect targets and perform semantic segmentation at the same time. In this paper, we propose a ship detection and segmentation method based on an improved Mask R-CNN model. We tested our method on the Airbus ship dataset and achieved very good results. We added a bottom-up structure and attention mechanisms to the Mask R-CNN. The main contributions of our paper are:

- Propose a method for ship detection and segmentation by enhancing the Mask R-CNN model. We added a bottom-up structure to the Mask R-CNN network and use channel-wise and spatial attention mechanisms to improve detection and segmentation accuracies.

II. RELATED WORK

We divide prior work on ship detection in satellite remote sensing images into the two main categories of traditional methods and deep learning methods.

1) **Traditional methods**. Traditional methods include statistical methods, transformation methods and others. Reference [28] is an example of statistical method, in which sea cluster histograms to construct anomaly detection models. It identifies candidate regions and then delete non-ship objects from the candidate regions. It uses structural continuity to remove false alarms. However, this method only performs well in the sea regions of satellite images from panchromatic and one band within multispectral and it does not perform well in the harbor area. Reference [29] proposed a model based on statistical analysis and shape identification. They perform statistical analysis on the ship distribution in the sea area to distinguish suspected ship targets from sea, land, islands or strong waves, and then use shape features such as aspect ratio, roundness, etc., to detect ships. In the transformation methods, a transform, e.g., Radon transform, wavelet transform or Hough transform is applied to the input image in order to extract features. For example, [30] proposes to use wavelet decomposition to obtain high- and low-frequency features of the image. The features are then combined by normalization and addition and a salient feature map is produced. In [31], a new ship detection method based on complex signal kurtosis (CSK) in single-channel SAR imagery was proposed. The method consists of two main parts: region proposal and target identification. They first detect potential ship locations based on the region proposal, and then identify the ships in these locations.

2) **Deep learning methods**. In recent years, many remote sensing satellite ship detection methods based on deep learning have been proposed. And the deep learning methods can be divided into two categories according to the marking methods used. In the first category, ships are labeled at the pixel level [32], [33]. Cheng *et al.* [32] proposed a FCN-based edge detection network. In the second category, ships are labeled by using a bounding box, e.g. in [17] and [20], Faster R-CNN was used to detect ships. Reference [21] proposes a densely connected multiscale neural network based on faster-RCNN for multiscale and multi-scene SAR ship detection. Reference [23] proposes a new network architecture by using squeeze and excitation mechanism in Faster R-CNN. Their proposed network suppresses redundant sub-features and further improve the ship detection rate in Sentinel-1 images. Reference [25] proposes the Rotation Dense Feature Pyramid Networks (R-DFPN) to detect ships from different scenes, including ocean and seaports. Their networks performed very well in remote sensing images from Google Earth. Reference [26] proposes a hierarchical selective filtering layer in their region proposal network to map features in different scales to the same scale space, which allows for efficient detection of ships at different scales. Their network is end-to-end and can detect inshore and offshore ships ranging in size from dozens to thousands of pixels.

Most of the methods referenced above can detect ships but they do not perform ship segmentation down to the pixel level. The existing ship detection methods using bounding box can also obtain substantial information of ships for analysis, but the bounding box also contains background pixels. In this paper, we propose an improved deep learning neural network based on Mask R-CNN [27] to simultaneously detect and segment ships in a single framework. By performing segmentation, the mask of the ship is obtained, which contains no background pixels. Shape features, such as area and perimeter of the ship, can be more accurately computed and used in the classification of the ship or for other analysis tasks.

A. MASK R-CNN

Mask R-CNN [27] is a simple and efficient instance segmentation model that can be used for tasks such as human pose recognition. It won first place in the COCO 2016 Challenge. Mask R-CNN combines Faster R-CNN [10] for target

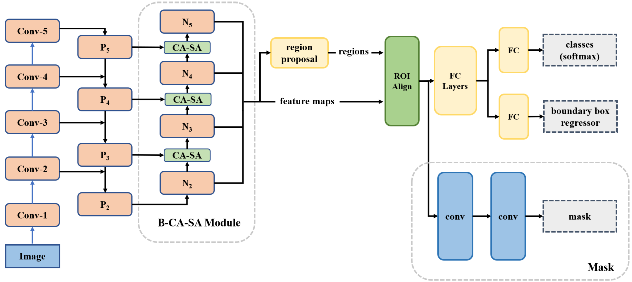


FIGURE 2. Illustration of the network: the B-CA-SA module represents the bottom-up structure with channel-wise and spatial attention mechanisms.

detection and FCN [34] for semantic segmentation. After the Faster R-CNN detects the target, FCN is used for mask prediction, border regression and classification. The effective combination of the two makes Mask R-CNN an excellent tool for object detection and segmentation.

B. ATTENTION MECHANISMS

The use of attention mechanism has produced very good results in many visual computing tasks, e.g., image classification [35] and pose estimation [36]. In [35], an attention residual learning mechanism was used to train the residual network for image classification. The SCA-CNN network proposed in [37] combines channel-wise and spatial-wise attention in CNN for image captioning. The SCRDet network proposed in [14] uses a pixel attention network and a channel attention network to suppress noise and highlight objects feature to detect small and cluttered objects. Reference [38] proposes a squeeze-and-excitation (SE) block to adaptively recalibrate channel-wise feature responses. This is achieved by explicitly modeling the interdependencies between channels. Their proposed network can bring significant improvements in the performance of current state-of-the-art CNNs. Inspired by these attention mechanisms, we use channel-wise attention and spatial attention in the bottom-up structure of our network. Experimental results showed that these attention mechanisms can improve detection and segmentation accuracies.

III. METHOD

Our proposed framework is illustrated in Figure 2. In the Mask R-CNN, we use FPN [39] to get the feature pyramids. We add a bottom-up path to shorten the information path between the lower layers and the topmost feature layer. This makes it easier for information in the lower layer to propagate to the top layer [40]. The channel-wise and spatial attention mechanisms used in the bottom-up path make the feature maps respond better to target's features.

A. BOTTOM-UP PATH STRUCTURE

The ResNet101 [41] network is commonly used in Mask R-CNN to extract features. In deep neural networks, the features in the lower layers pass through dozens of network layers to reach the top layers. After passing through many layers, some of the lower-level information may be lost. However, information contained in the lower level features are important for instance segmentation. In PANet [40], Liu et al. proposed a bottom-up path argumentation technique to shorten the information path and enhance the feature pyramid with accurate localization signals from the lower levels. Inspired by PANet [40], we have adopted a similar approach. We use ResNet and FPN structures to get the feature maps of four feature levels, namely $\{P_2, P_3, P_4, P_5\}$. As shown in Figure 3, the bottom-up augmented path goes from level P_2 to level P_5, and the size of the feature maps at each feature level N_i (i = 2, 3, 4, and 5) is the same as that of the corresponding level P_i (i = 2, 3, 4, and 5). This is then sent to the new feature maps $\{N_2, N_3, N_4, N_5\}$ to the subsequent network layers instead of $\{P_2, P_3, P_4, P_5\}$. Its network structure is shown in Figure 3.

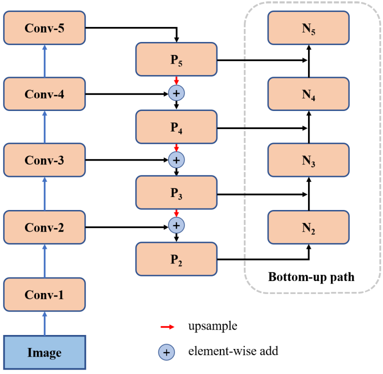


FIGURE 3. FPN with bottom-up path structure from N_2 to N_5. On the left is FPN structure without any changes.

B. CHANNEL-WISE AND SPATIAL ATTENTION MECHANISMS

1) CHANNEL-WISE ATTENTION

Generally, features obtained at different channels in CNN contain different semantic information. The features obtained from different channels are not all equal in terms of importance. Some channels may not contain any target feature. As shown in Figure 4, which contains the P_3 feature maps obtained by convolution, different channels contain different information. Channels 167 and 85 contain useful ship information but channel 97 contains information mostly for the background. When extracting features by convolution, most existing methods assign the same weight to the different channels and do not carry out channel selection. In our channel-wise attention mechanism, the channels with higher target responses are allocated larger weights. This allows us to get the desired object features.

For the channel-wise attention, we have the convolution features $F = [F_1, F_2, \ldots, F_C]$, where $F_i \in R^{w \times h}$ denotes the i-th channel of the feature map F, and C is the total number of channels. We apply the average pooling operation to each feature map F_i and produce a channel feature vector V:

$$V = [V_1, V_2, \ldots, V_c], \quad V \in R^c$$

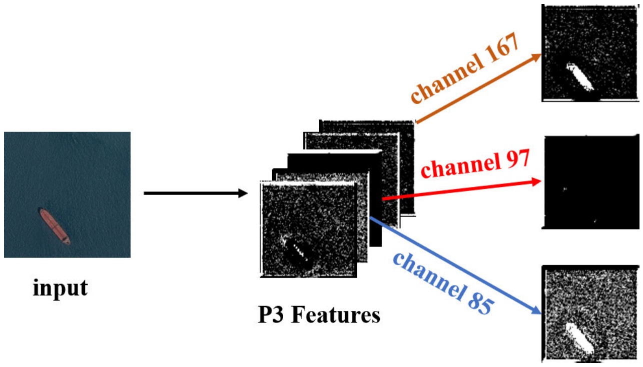


FIGURE 4. Sample features in different channels.

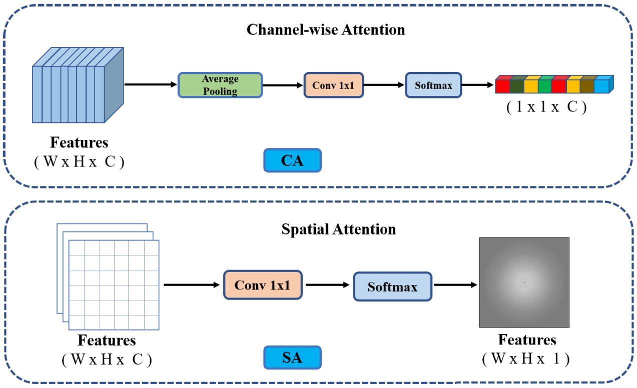


FIGURE 5. Illustration of Channel-wise Attention (CA) and Spatial Attention (SA).

where V_i represents the feature from the i-th channel after average pooling. Convolution with a 1×1 kernel is then performed to learn the aggregated features from each channel V_i. The softmax operation is then performed on the feature vector V so that the sum from all channels is 1.

2) SPATIAL ATTENTION

In object detection from images, the objects we would like to detect appear in some parts of the image but not the whole image. Different from the regular CNN network, which treats each region in the image equally, the spatial attention mechanism assigns a weight to each pixel in the feature map. This allows more attention to be paid to pixels that belong to the foreground region. Many studies have proved the effectiveness of the spatial attention mechanism in its ability to reduce background interferences [42], [43]. Our spatial attention mechanism works as follows. Given the convolution features $F = [F_1, F_2, \ldots, F_C]$, we use a convolution kernel of size 1×1 to generate a feature map summary $\mathbf{M}$. The softmax operation is then performed on the pixel points of the feature $\mathbf{M}$ so that they sum up to one.

C. FPN AND BOTTOM-UP STRUCTURES WITH SPATIAL AND CHANNEL-WISE ATTENTION MECHANISMS

In our experiments, we tried different combinations of the spatial attention and channel-wise attention mechanisms with the FPN and bottom-up structures.

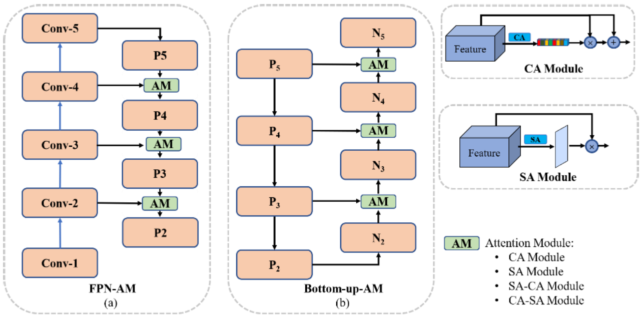


FIGURE 6. (a) FPN with attention module. (b) Bottom-up structure with attention module. Four different types of Attention Modules (AM): Channel-wise Attention (CA) Module, Spatial Attention (SA) Module, SA-CA module (combination of SA and CA modules), and CA-SA module (combination of CA and SA modules).

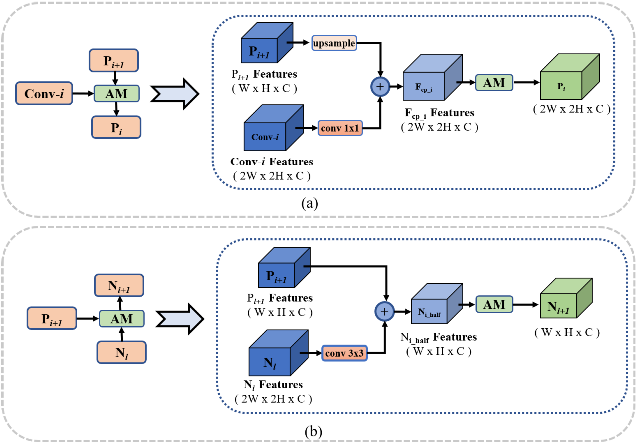


FIGURE 7. (a) Illustration of the building blocks of FPN-AM. (b) Illustration of the building blocks of B-AM.

1) FPN WITH ATTENTION MODULE (FPN-AM)

We first obtain four levels of feature maps by using ResNet: $\{C_2, C_3, C_4, C_5\}$. To generate the feature pyramid $\{P_2, P_3, P_4, P_5\}$, we combine the FPN with attention module through the following steps. P_5 is generated by feature map C_5 by using a convolution kernel of size 1×1 and 256 channels. For the other levels, as shown in Figure 7(a), we up sample the feature map P_{i+1} (i = 4, 3, and 2), by using a up sample factor of 2, then the up sampled feature map is merged with corresponding feature map C_i (a 1×1 convolutional layer is used to reduce the channel dimensions to 256) by element-wise addition to generate a new feature map F_{cp_i}. Feature map F_{cp_i} is then send to the selected attention module (CA, SA, CA-SA or SA-CA) to obtain feature map P_i (i = 4, 3, and 2). Finally, we obtain the feature pyramid $\{P_2, P_3, P_4, P_5\}$ from the FPN-AM structure.

2) BOTTOM-UP WITH ATTENTION MODULE (BOTTOM-UP-AM)

This structure is similar to FPN-AM. As shown in Figure 6 (b). N_2 is the same as P_2. To generate the feature map N_{i+1} (i = 2, 3, and 4), as shown in Figure 7(b),

N_i (2W×2H×C) is convolved by a convolution kernel of size 3 × 3, with a stride size of 2. The height and width of the feature maps are reduced by one-half so that they are of the same size as P_{i+1} (W×H×C). The reduced feature maps are merged with P_{i+1} by element-wise addition to get a new feature map N_{i_half}, which is then sent to the attention module (CA, SA, CA-SA or SA-CA) to get feature map N_i. Finally, a new feature pyramid $\{N_2, N_3, N_4, N_5\}$ is obtained and then sent to subsequent network layers.

IV. EXPERIMENTAL RESULTS

We used the dataset for airbus ship detection challenge [46] in our experiments. Reference [44] improves Mask R-CNN by replacing NMS in Mask R-CNN with Soft-NMS and achieved good results in their ship detection and segmentation experiments. We have implemented the network in [44] and ran it on the airbus dataset. Mask Scoring R-CNN [45] is an improved variant of Mask R-CNN obtained by adding to Mask R-CNN a network block that will learn the quality of the predicted instance masks. We also ran the Mask Scoring R-CNN network on the airbus dataset. The bottom-up structure in our network was inspired by PANet [40], which is also an improved variant of mask R-CNN. We ran PANet on the airbus ship dataset and compared the results with ours. SCRDet [14] proposed a rotating bounding box method to detect rotated objects in remote sensing images. We also ran SCRDet on the airbus ship dataset for comparison. In addition, we also performed ablation studies to verify the effectiveness of the proposed channel-attention mechanism, spatial attention and the bottom-up path structure. Experimental results show that our proposed method performs better than baseline Mask R-CNN, Mask Scoring R-CNN [45], PANet [40], SCRDet [14], and the method in [44].

A. DATASET

As there is a lack of datasets with ground truths for ship segmentation, Airbus created a large dataset with masks for the ship regions. They also held a competition on Kaggle [46]. The Airbus ship dataset consists of remote sensing images of ships with sea and harbor in the background. It contains training and test images with masks for the ships. We selected and used 42, 500 images from the Airbus ship dataset in our experiments. For testing, we randomly selected 3, 000 images from our dataset and used the remaining images for training. The original dataset was encoded in RLE (run-length encoding) format. To facilitate training, we converted it into COCO annotation format. In order to compare with SCRDet [14], we also used horizontal bounding boxes and oriented bounding boxes to annotate the dataset.

B. IMPLEMENTATION DETAILS

The Pytorch framework was used in our experiments and the basic code used was Facebook Research's Mask R-CNN Benchmark [47]. In our experiments, we use the pre-trained ResNet-101 model for initialization. The GPU used in our experiment was a GTX 1080 with 8 GB memory. The training and testing images are of size 768×768. For data augmentation, we performed random horizontal flips of the training images. During training, the batch size was set to 1, initial learning rate set to 0.001, weight attenuation set to 0.0001 and synchronized SGD with a momentum of 0.9 was used as optimizer. The maximum number of iterations was set to 350, 000. In the RPN network, we assign a single scale anchor point at each level, and we assigned five scales $\{32^2, 64^2, 128^2, 256^2, 512^2\}$ anchor points at each level, respectively $\{N_2, N_3, N_4, N_5, N_6\}$ (N_6 is obtained by performing max pooling on N_5), and the anchor points at each level have the aspect ratio $\{1:1, 1:2, 2:1\}$.

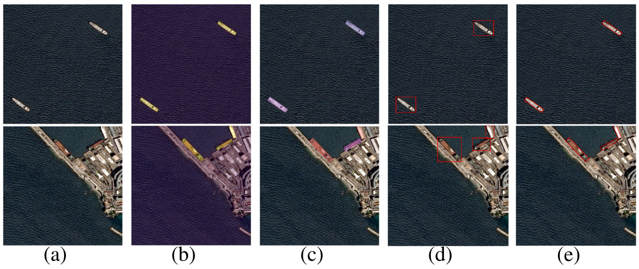


FIGURE 8. Translation of annotation format: (a) original image, (b) ground truths in RLE format, (c) ground truths in COCO annotation format, (d) ground truths in horizontal bounding box format, (e) ground truths in oriented bounding box format.

C. EVALUATION AND RESULTS

We used ResNet101 as the backbone network to extract features and compare with the Mask R-CNN baseline model, PANet [40], Mask Scoring R-CNN [45], method in [44], and our methods trained models. We use the standard metrics average precision (AP, AP_{50}, AP_{75}, AP_S, AP_M, AP_L) to evaluate our results. Experimental results are shown in Table 1, 2, 3, and 4.

1) THE EFFECTS OF SPATIAL AND CHANNEL-WISE ATTENTION

We choose Mask R-CNN as the baseline in our ablation study. For fair comparisons, all experimental data and parameter settings are kept the same.

a: EFFECTS OF SPATIAL ATTENTION

Tables 1 and 2 show that the spatial attention structure is beneficial in the suppression of noise and in highlighting object information. With the addition of spatial attention to the FPN structure, ship detection and segmentation accuracies improved by 4.9% and 3.2%, respectively. In addition, by adding spatial attention to the bottom-up structure, the accuracies have improved by 2.3% and 1.9%, respectively.

b: EFFECTS OF CHANNEL-WISE ATTENTION

Tables 1 and 2 show that the channel-wise attention structure can assign larger weights to channels which show higher responses to objects and alleviate the influence of the

TABLE 1. Detection accuracy in the ablation study of spatial attention, channel-wise attention and bottom-up structure. SA, CA and B stand for spatial attention, channel-wise attention and bottom-up structure, respectively.

	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN baseline	70.6	95.4	79.9	62.0	86.4	86.9
Mask R-CNN FPN-SA	75.5	**96.0**	85.2	68.3	89.2	90.5
Mask R-CNN FPN-CA	74.9	95.9	84.2	67.6	88.6	89.9
Mask R-CNN FPN-SA-CA	75.1	95.9	84.3	67.6	89.4	**91.1**
Mask R-CNN FPN-CA-SA	75.3	95.9	84.9	67.7	**89.5**	91.0
Mask R-CNN B	73.6	95.6	84.0	67.6	86.7	77.2
Mask R-CNN B-SA	75.9	**96.0**	85.5	68.8	**89.5**	89.9
Mask R-CNN B-CA	76.0	95.9	85.4	**69.1**	89.3	88.8
Mask R-CNN B-SA-CA	75.8	**96.0**	85.3	68.8	89.1	89.5
Mask R-CNN B-CA-SA	**76.1**	95.9	**86.4**	**69.1**	**89.5**	89.5

TABLE 2. Segmentation accuracy in the ablation study of spatial attention, channel-wise attention and bottom-up structure.

	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN baseline	62.0	92.3	70.6	50.6	79.9	84.6
Mask R-CNN FPN-SA	65.2	93.6	74.1	54.1	82.6	86.7
Mask R-CNN FPN-CA	64.9	93.5	72.9	53.8	82.4	87.1
Mask R-CNN FPN-SA-CA	64.7	93.5	73.3	53.3	82.8	87.3
Mask R-CNN FPN-CA-SA	65.0	93.5	73.1	53.3	**83.0**	**87.4**
Mask R-CNN B	63.7	93.1	72.7	54.1	80.0	78.3
Mask R-CNN B-SA	65.6	**93.8**	74.4	**55.2**	82.5	85.9
Mask R-CNN B-CA	65.7	**93.8**	74.5	55.1	82.8	86.0
Mask R-CNN B-SA-CA	**65.8**	93.6	74.5	55.1	82.8	85.6
Mask R-CNN B-CA-SA	**65.8**	93.6	**75.5**	55.1	82.8	85.3

TABLE 3. Detection accuracy of different methods.

	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN baseline	70.6	95.4	79.9	62.0	86.4	86.9
Mask R-CNN + S-NMS [44]	66.5	88.6	75.3	55.4	85.7	88.3
Mask Scoring R-CNN [45]	65.8	93.5	72.5	54.5	85.3	88.7
PANet [40]	72.8	95.1	80.4	63.2	**89.7**	**91.3**
Mask R-CNN B-CA-SA (ours)	**76.1**	**95.9**	**86.4**	**69.1**	89.5	89.5

TABLE 4. Segmentation accuracy of different methods.

	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN baseline	62.0	92.3	70.6	50.6	79.9	84.6
Mask R-CNN + S-NMS [44]	58.6	86.4	67.6	46.5	79.4	85.2
Mask Scoring R-CNN [45]	56.2	89.7	60.2	42.0	78.1	85.2
PANet [40]	62.0	82.3	68.5	49.4	79.5	84.6
Mask R-CNN B-CA-SA (ours)	**65.8**	**93.6**	**75.5**	**55.1**	**82.8**	**85.3**

background. With the addition of channel-wise attention to the FPN structure, ship detection and segmentation accuracies improved by 4.3% and 2.9%, respectively. In addition, by adding channel-wise attention to the bottom-up structure, ship detection and segmentation accuracies improved by 2.4% and 2.0%, respectively.

When spatial attention is combined with channel-wise attention, detection and segmentation accuracies greatly improved over those of the baseline. When compared to spatial attention alone or channel-wise attention alone, combining the two caused slight increase or slight decrease in accuracies.

TABLE 5. Detection accuracy (or score) of different tasks.

	HBB mAP[0.05:0.95]	HBB mAP@0.5	OBB mAP@0.5	Kaggle leaderboard
Mask R-CNN baseline	70.6	95.4	—	83.62
Mask R-CNN + S-NMS [44]	66.5	88.6	—	83.67
Mask Scoring R-CNN [45]	65.8	93.5	—	84.04
PANet [40]	72.8	95.1	—	82.52
Mask R-CNN B-CA-SA (ours)	**76.1**	**95.9**	—	84.11
SCRDet [14]	—	74.5	40.6	75.60
Kaggle Winner [46]	—	—	—	**85.45**

2) THE EFFECTS OF BOTTOM-UP STRUCTURE

As shown in Tables 1 and 2, when the bottom-up structure is added to Mask R-CNN, the detection and segmentation accuracies for small ships improved by 5.6% and 3.5%, respectively. The improvements are likely due to the added bottom-up structure which shortens the information path and enhances the feature pyramid with accurate localization signals from the lower levels. However, detection and segmentation accuracies for large ships decreased by 9.7% and 6.3%, respectively. This may be due to the reduction of top-level feature information caused by the addition of the bottom-up structure. But, with the addition of bottom-up structure, ship detection and segmentation mAPs increased by 3.0% and 1.7% respectively. In addition, we found that B-AM, which uses bottom-up structure, has slightly better detection and segmentation accuracies than FPN-AM when the same attention structure is used.

3) COMPARISONS WITH OTHER METHODS

In this section, we compare our proposed method with the Mask R-CNN baseline model, PANet [40], Mask Scoring R-CNN [45], SCRDet [14], and the method in [44]. Comparisons of detection and segmentation accuracies are shown in Table 3, 4, and 5.

Reference [44] improved the Mask R-CNN network by replacing NMS with Soft-NMS and then use the improved network for the detection and segmentation of inshore ships. Their improved network was able to increase the detection and segmentation accuracies for large-sized ships but accuracies for small-sized ships were greatly reduced. The detection and segmentation accuracies for small-sized ships were 6.6% and 4.1% lower than those of the Mask R-CNN baseline model, respectively. The reason for lower accuracies may be due to that Soft-NMS was proposed mainly for the detection of overlapping objects in an image. However, the ships in our data set are basically isolated objects and Soft-NMS may be less effective than regular NMS. Mask Scoring R-CNN [45] is an improved variant of Mask R-CNN and it achieved better performance on the COCO dataset for instance segmentation but it did not work well on our ship dataset. It performed worse than the baseline Mask R-CNN network except for large-sized ships. PANet [40] improved Mask R-CNN by adding a bottom-up path and an adaptive feature pooling mechanism. In our experiments, PANet improved ship detection accuracy by 2.2%, especially for medium- and large-sized ships, but ship segmentation accuracy was about the same as baseline Mask R-CNN. Our proposed network achieved the largest improvement in detection and segmentation accuracies: by 5.5% and 3.8%, respectively. We obtained significant improvements in accuracies for ships of all sizes (small, medium and large) and especially for small-sized ships. The effectiveness of our proposed network has therefore been demonstrated.

As shown in Figure 8 (d) and (e), we converted the annotation format of the ship data set from rle-mask to bounding box. Then we train the SCRDet [14] network on the converted dataset. We used a weight decay value of 0.0001 and the momentum used was 0.9. We trained for a total of 350,000 iterations. The learning rate changed from 3e-4 to 3e-6 between the $100,000^{th}$ and $200,000^{th}$ iterations. We used mAP@0.5 to evaluate the detection results from the horizontal bounding box (HBB) and oriented bounding box (OBB). As shown in Table 5, the accuracy for OBB is 40.6% and the accuracy for HBB is 74.5%, which are lower than that of our model with an accuracy of 95.9% (mAP@0.5).

To compare with other methods, we tested our models on the original Airbus Ship Dataset (with 15,606 images) and submitted the results to the Kaggle leaderboard. As shown in Table 5, our model obtained a score of 84.11. Compared with the Mask R-CNN baseline, Mask Scoring R-CNN, PANet and the method in [44], our score is slightly better. Our score is slightly lower than the Kaggle winner's score of 85.45. Many participants in the competition, including the winner, did not publish a paper or provide a detailed description of the method they used. This makes it difficult to compare their methods with ours other than comparing the scores. However, through the discussions they had among the contest participants in the discussion forum, many participants used multiple networks and then combine the results in order to boost their score on the leaderboard. In contrast, we used a single end-to-end network, which is a better approach in practical applications.

D. DISCUSSION

Experiments on the expanded Airbus ship dataset have verified the effectiveness of our proposed network for ship detection and segmentation in satellite remote sensing images. We demonstrated that our network is capable of suppressing

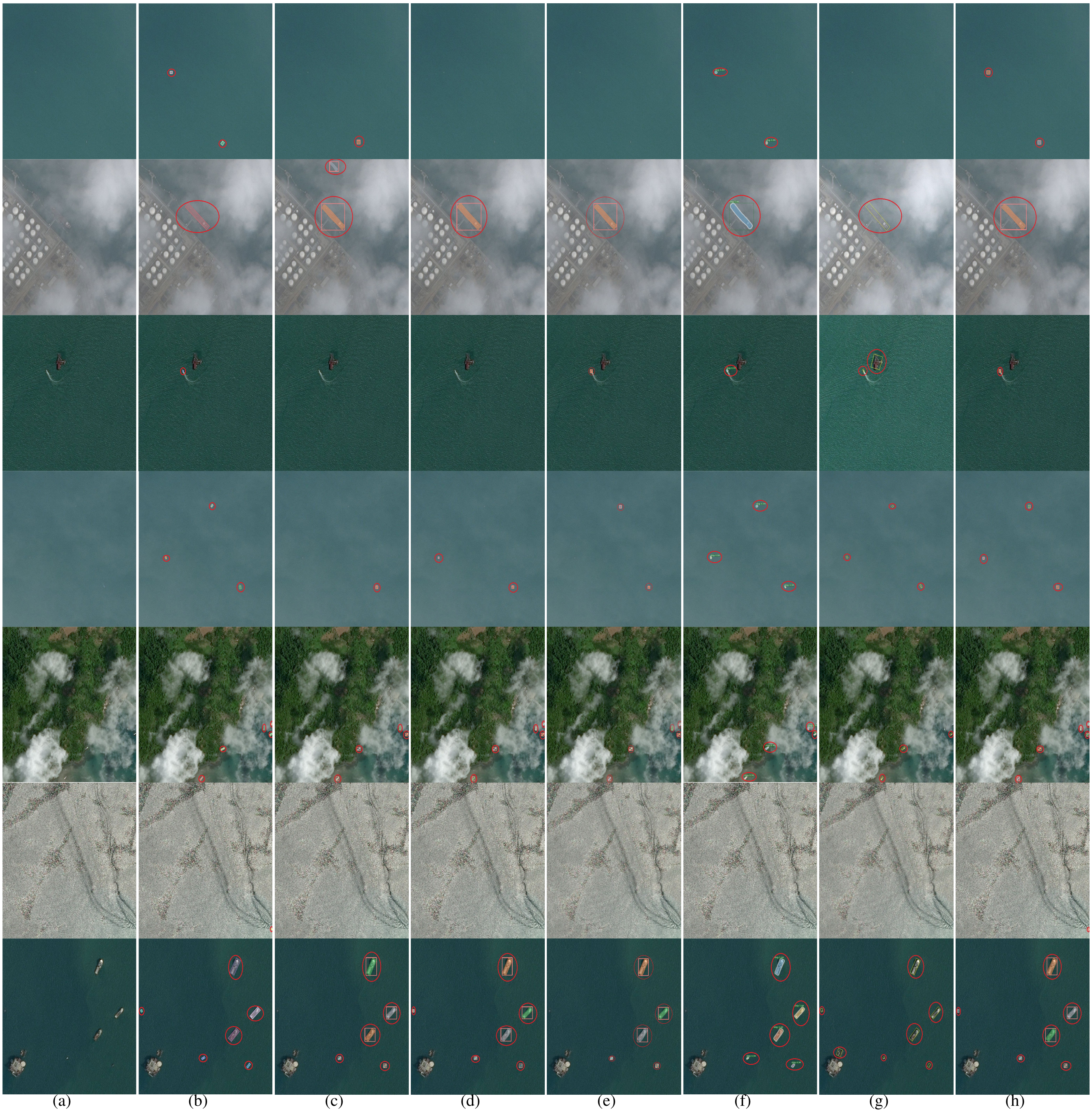


FIGURE 9. Samples of ships detection and segmentation: (a) original images, (b) ground truths, (c) results from Mask R-CNN baseline model, (d) results from Mask R-CNN [44], (e) results from Mask Scoring R-CNN model, (f) results from PANet model, (g) results from SCRDet model, (h) results from Mask R-CNN_B-CA-SA (ours). Ships in the ground truths and objects detected by the models are circled in red.

noise and highlighting object information. This is achieved by assigning a corresponding weight value to each pixel in the feature map in the spatial attention mechanism. The channel attention mechanism assigns large weights to the channels with high target response, thus avoiding interference from the background. Both the channel attention and spatial attention mechanisms are helpful in improving the ship detection and segmentation accuracies. By adding a bottom-up structure to the FPN structure of Mask R-CNN, the path between the lower layers and the top-most layer is shortened. This allows lower layer features to be more effectively utilized at the top layer. From our experiments, we also found that the spatial attention mechanism and the channel attention mechanism can significantly improve the detection and segmentation accuracies when used alone. However, compared to spatial attention alone or channel-wise attention alone, combining the two caused a slight decrease in accuracies. We suspect that combining the two attention mechanisms may cause some degree of information loss in small targets, and therefore combining the two does not increase the accuracies. We will investigate this further in future work.

V. CONCLUSION

We have proposed an end-to-end deep learning network for ship detection and segmentation in remote sensing satellite images. Our network is constructed by adding attention mechanisms and a bottom-up structure to the Mask R-CNN deep learning network. Adding the attention mechanisms and the bottom-up structure enhances the propagation of information from lower-layers to the top layers. This significantly improved the overall detection and segmentation accuracies when compared to the baseline model and other methods. We also observed from our experiments that the detection and segmentation accuracies of our method for small ships can be further improved. We will address this in our future work.

REFERENCES

[1] J. Huang, Z. Jiang, H. Zhang, B. Cai, and Y. Yao, "Region proposal for ship detection based on structured forests edge method," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 1856–1859.

[2] C. Wang, F. Bi, W. Zhang, and L. Chen, "An intensity-space domain CFAR method for ship detection in HR SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 529–533, Apr. 2017.

[3] X. Leng, K. Ji, K. Yang, and H. Zou, "A bilateral CFAR algorithm for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 7, pp. 1536–1540, Jul. 2015.

[4] X. Leng, K. Ji, X. Xing, S. Zhou, and H. Zou, "Area ratio invariant feature group for ship detection in SAR imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2376–2388, Jul. 2018.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[6] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lille, France, Jul. 2015, pp. 448–456.

[7] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 5353–5360.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 580–587.

[9] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1440–1448.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[11] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Banff, AB, Canada, Apr. 2014.

[12] J. Redmon and A. Farhadi, "YOLOV3: An incremental improvement," *CoRR*, vol. abs/1804.02767, pp. 1–6, Apr. 2018.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 21–37.

[14] X. Yang, K. Fu, H. Sun, J. Yang, Z. Guo, M. Yan, T. Zhang, and X. Sun, "R2CNN++: Multi-dimensional attention based rotation invariant detector with robust anchor strategy," *CoRR*, vol. abs/1811.07126, pp. 1–10, 2018.

[15] J. Ding, N. Xue, Y. Long, G. Xia, and Q. Lu, "Learning RoI transformer for detecting oriented objects in aerial images," *CoRR*, vol. abs/1812.00155, pp. 1–19, 2018.

[16] Q. Li, L. Mou, Q. Xu, Y. Zhang, and X. X. Zhu, "R^3-net: A deep network for multi-oriented vehicle detection in aerial images and videos," *CoRR*, vol. abs/1808.05560, pp. 1–14, 2018.

[17] M. Kang, X. Leng, Z. Lin, and K. Ji, "A modified faster R-CNN based on CFAR algorithm for SAR ship detection," in *Proc. Int. Workshop Remote Sens. Intell. Process. (RSIP)*, May 2017.

[18] M. Kang, K. Ji, X. Leng, and Z. Lin, "Contextual region-based convolutional neural network with multilayer fusion for SAR ship detection," *Remote Sens.*, vol. 9, no. 8, p. 860, Aug. 2017.

[19] S. Wang, G. Xin, S. Hao, X. Zheng, and S. Xian, "An aircraft detection method based on convolutional neural networks in high-resolution SAR images," *J. Radars*, to be published.

[20] J. Li, C. Qu, and J. Shao, "Ship detection in SAR images based on an improved faster R-CNN," in *Proc. SAR Big Data Era, Models, Methods Appl. (BIGSARDATA)*, Nov. 2017.

[21] J. Jiao, Y. Zhang, H. Sun, X. Yang, X. Gao, W. Hong, K. Fu, and X. Sun, "A densely connected end-to-end neural network for multiscale and multiscene SAR ship detection," *IEEE Access*, vol. 6, pp. 20881–20892, 2018.

[22] Z. Cui, S. Dang, Z. Cao, S. Wang, and N. Liu, "SAR target recognition in large scene images via region-based convolutional neural networks," *Remote Sens.*, vol. 10, no. 5, p. 776, May 2018.

[23] Z. Lin, K. Ji, X. Leng, and G. Kuang, "Squeeze and excitation rank faster R-CNN for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 751–755, May 2019.

[24] Y.-L. Chang, A. Anagaw, L. Chang, Y. Wang, C.-Y. Hsiao, and W.-H. Lee, "Ship detection based on YOLOv2 for SAR imagery," *Remote Sens.*, vol. 11, no. 7, p. 786, Apr. 2019.

[25] X. Yang, H. Sun, K. Fu, J. Yang, X. Sun, M. Yan, and Z. Guo, "Automatic ship detection of remote sensing images from Google earth in complex scenes based on multi-scale rotation dense feature pyramid networks," *CoRR*, vol. abs/1806.04331, 2018.

[26] Q. Li, L. Mou, Q. Liu, Y. Wang, and X. X. Zhu, "HSF-Net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7147–7161, Dec. 2018.

[27] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *CoRR*, vol. abs/1703.06870, 2017.

[28] X. Xiaoyang, Q. Xu, and H. Lei, "Fast ship detection from optical satellite images based on ship distribution probability analysis," in *Proc. 4th Int. Workshop Earth Observ. Remote Sens. Appl. (EORSA)*, Jul. 2016.

[29] Y. Ji-Yang, H. Dan, W. Lu-Yuan, G. Jian, and W. Yan-Hua, "A real-time on-board ship targets detection method for optical remote sensing satellite," in *Proc. IEEE 13th Int. Conf. Signal Process. (ICSP)*, Nov. 2016.

[30] Y. Ji-Yang, H. Dan, W. Lu-Yuan, L. Xin, and L. Wen-Juan, "On-board ship targets detection method based on multi-scale salience enhancement for remote sensing image," in *Proc. IEEE 13th Int. Conf. Signal Process. (ICSP)*, Nov. 2016.

[31] X. Leng, K. Ji, S. Zhou, and X. Xing, "Ship detection based on complex signal kurtosis in single-channel SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6447–6461, Sep. 2019.

[32] D. Cheng, G. Meng, S. Xiang, and C. Pan, "FusionNet: Edge aware deep convolutional networks for semantic segmentation of remote sensing harbor images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 12, pp. 5769–5783, Dec. 2017.

[33] H. Lin, Z. Shi, and Z. Zou, "Fully convolutional network with task partitioning for inshore ship detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1665–1669, Oct. 2017.

[34] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[35] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6450–6458.

[36] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5669–5678.

[37] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6298–6306.

[38] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.

[39] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 936–944.

[40] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake, UT, USA, Jun. 2018, pp. 8759–8768.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[42] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 3–19.

[43] Z. Laskar and J. Kannala, "Context aware query image representation for particular object retrieval," in *Proc. 20th Scandin. Conf. Image Anal. (SCIA)*, Tromsø, Norway, Jun. 2017, pp. 88–99.

[44] S. Nie, Z. Jiang, H. Zhang, B. Cai, and Y. Yao, "Inshore ship detection based on mask R-CNN," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018.

[45] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring R-CNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6409–6418.

[46] Airbus. *Airbus Ship Detection Challenge*. Accessed Feb. 14, 2019. [Online]. Available: https://www.kaggle.com/c/airbus-ship-detection

[47] F. Massa and R. Girshick. (2018). *maskrCNN-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch*. Accessed: Feb. 14, 2019. [Online]. Available: https://github.com/facebookresearch/maskrcnn-benchmark

XUAN NIE was born in 1976. He received the B.S. degree in automatic control, the M.S. degree in pattern recognition, and the Ph.D. degree in computer application technology from Northwestern Polytechnical University, Xi'an, China, in 1998, 2001, and 2005, respectively. In 2006, he joined as a Lecturer with the School of Software, Northwestern Polytechnical University, where he has been an Associate Professor, since 2010. He was a Visiting Professor with The Hong Kong Polytechnical University, in 2010, and the University of Michigan, USA, in 2013. He is currently an Associate Professor with the School of Software, Northwestern Polytechnical University. He has authored or coauthored over 30 journal articles and conference papers, three monographs, and co-invented patents. His main research interests include machine learning, computer vision, image processing, and artificial intelligence. He was a Reviewer for the IEEE Internet of Things Journal. He is the Honor of the Science and Technology Achievement Award of Xi'an, in 2015.

MENGYANG DUAN was born in 1995. He received the B.S. degree in software engineering from Shandong University, Weihai, China, in 2018. He is currently pursuing the M.S. degree with Northwestern Polytechnical University, Xi'an, China. His current research interests include object segmentation and object detection.

HAOXUAN DING was born in 1995. He received the B.S. degree in flight vehicle propulsion engineering from Northwestern Polytechnical University, Xi'an, China, in 2018, where he is currently pursuing the M.S. degree. His current research interests include generative adversarial networks and object detection.

BINGLIANG HU received the M.S. and Ph.D. degrees from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences. He is currently the Associate Director of the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences. His research interests include the photoelectric signal processing and spectral imaging technology. He has published extensively in these areas. He also serves as a member of the Chinese Society of Image and Graphics, the Optical Engineering Society, the Third Xi'an Youth Science and Technology Association. He has presided over and completed 18 major national scientific research plans.

EDWARD K. WONG received the B.E. degree from the State University of New York at Stony Brook, the M.Sc. degree from Brown University, and the Ph.D. degree from Purdue University, all in electrical engineering. He is currently an Associate Professor with the Department of Computer Science and Engineering, NYU Tandon School of Engineering, Brooklyn, NY, USA. His research interests include the areas of computer vision, multimedia computing, medical image processing, and digital forensics. He has published extensively in these areas. He has worked on many research projects funded by federal and state agencies, as well as private industry. He has served as an Associate Editor for the Journal *Information Sciences* and the *International Journal of Multimedia Intelligence and Security*. He is also an Associate Editor for the Journal *LNCS Transactions on Data Hiding and Multimedia Security* (Springer). He has also served on the organizing committee and technical program committee of numerous IEEE, ACM, and other international conferences.